

Как мы делали отказоустойчивый Redis

Евгений Дюков,
разработчик Managed Databases



Поговорим про:

- Redis
- Sentinel
- Redis Cluster
- Network partitioning

Не поговорим про:

- KeyDB
- Redisraft
- Dragonfly DB
- Disk I/O failures

План доклада

1. Зачем в Redis отказоустойчивость?
2. Sentinel
3. Redis Cluster
4. Как с ними потерять данные
5. Компенсируем проблемы «дёшево и сердито»
6. Альтернативный подход
7. Замена Sentinel
8. Разбираем failover/switchover
9. Тесты

Redis
и отказоустойчивость

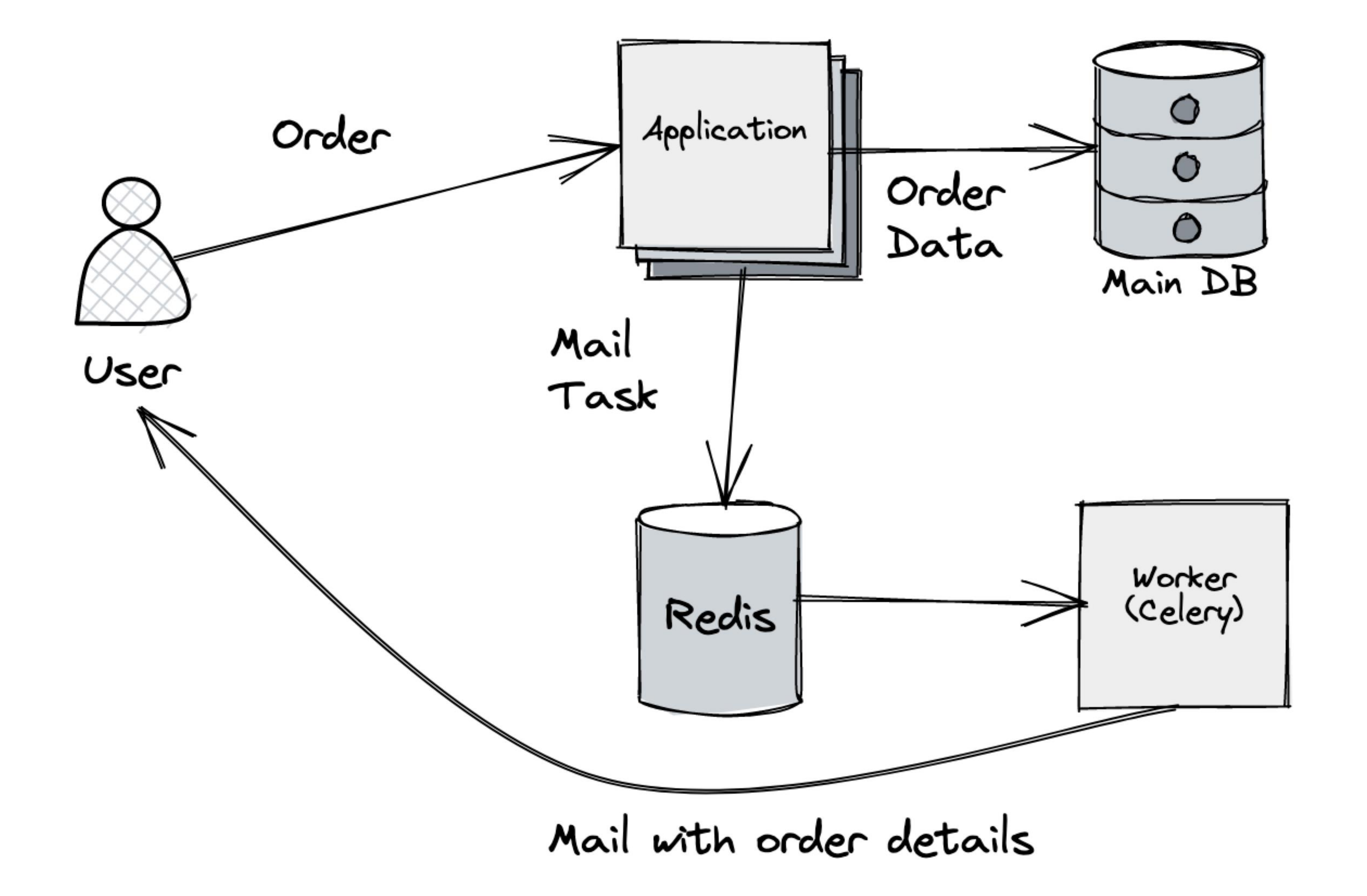
- «Redis — это только кэш»

Redis
и отказоустойчивость

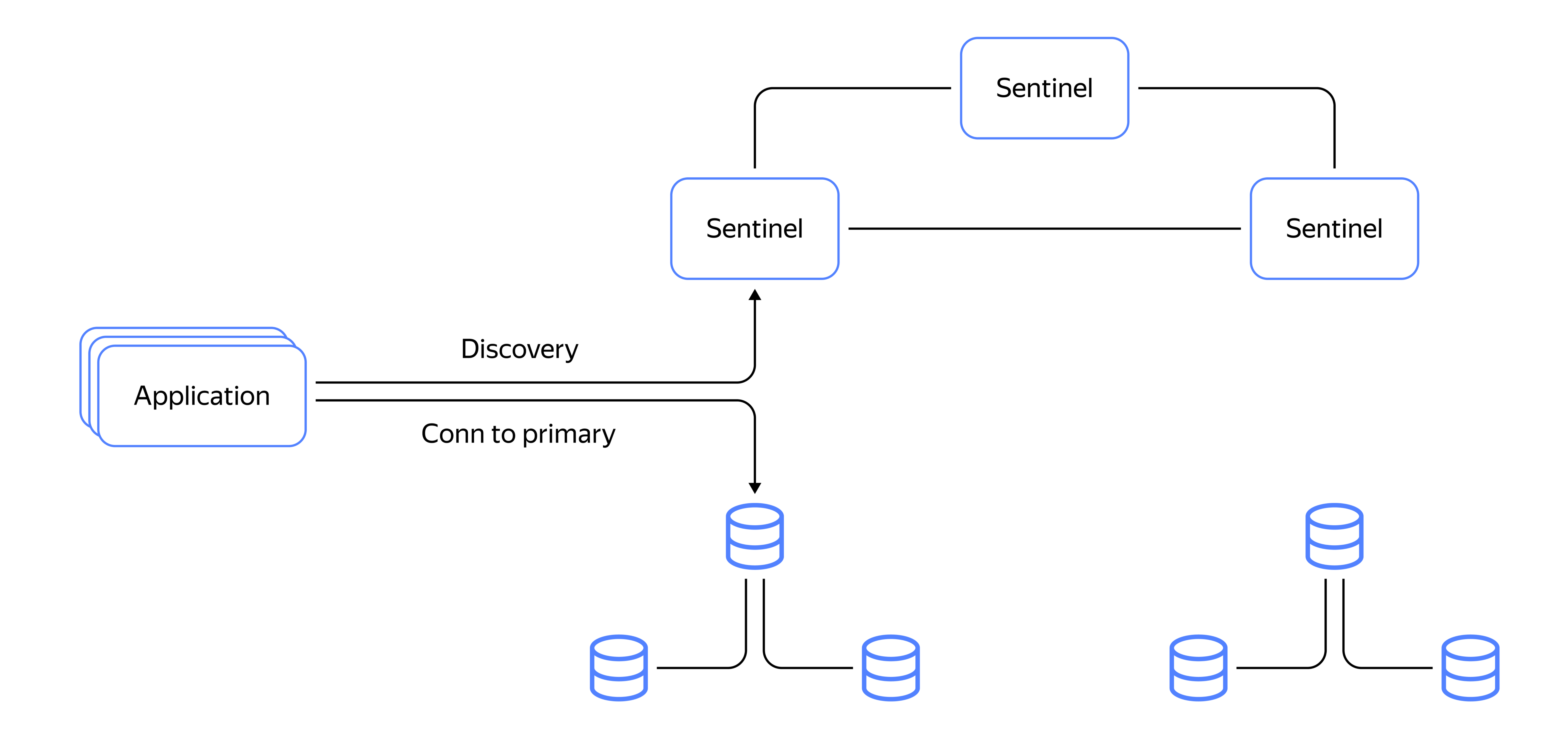
- «Redis — это только кэш»
- Очереди
- Пользовательские сессии
- Rate limiting

и так далее

Очереди?


Order
Application
Order
Data
Main DB
User
Mail
Task
Redis
Worker
(Celery)
Mail with order details

Отказоустойчивость в open source Redis: Sentinel

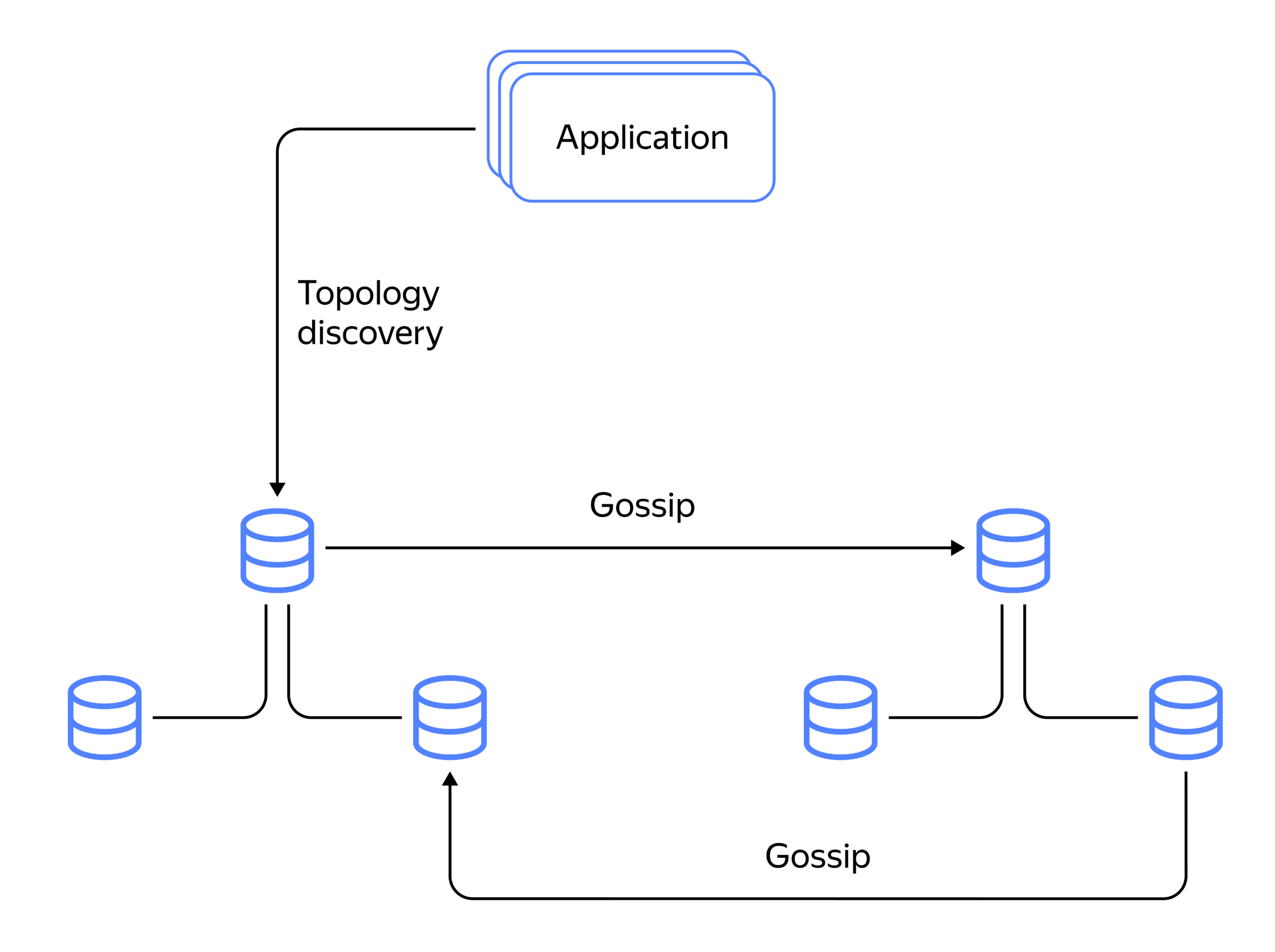


Фичи Sentinel

- Обнаружение реплик
- Способность обслуживать множество групп Primary + Replicas
- Pub/Sub для топологии

  Подписаться на новые реплики / изменение Primary
- Тот же протокол RESP2/RESP3
- Хорошо поддержан в клиентских библиотеках

Отказоустойчивость в open source Redis: Cluster



Фичи Redis Cluster

- Нет необходимости держать рядом отдельную сущность
- Переживает даже отказ majority в одном шарде
- Работа с кластером с точки зрения клиента проще

Проблемы

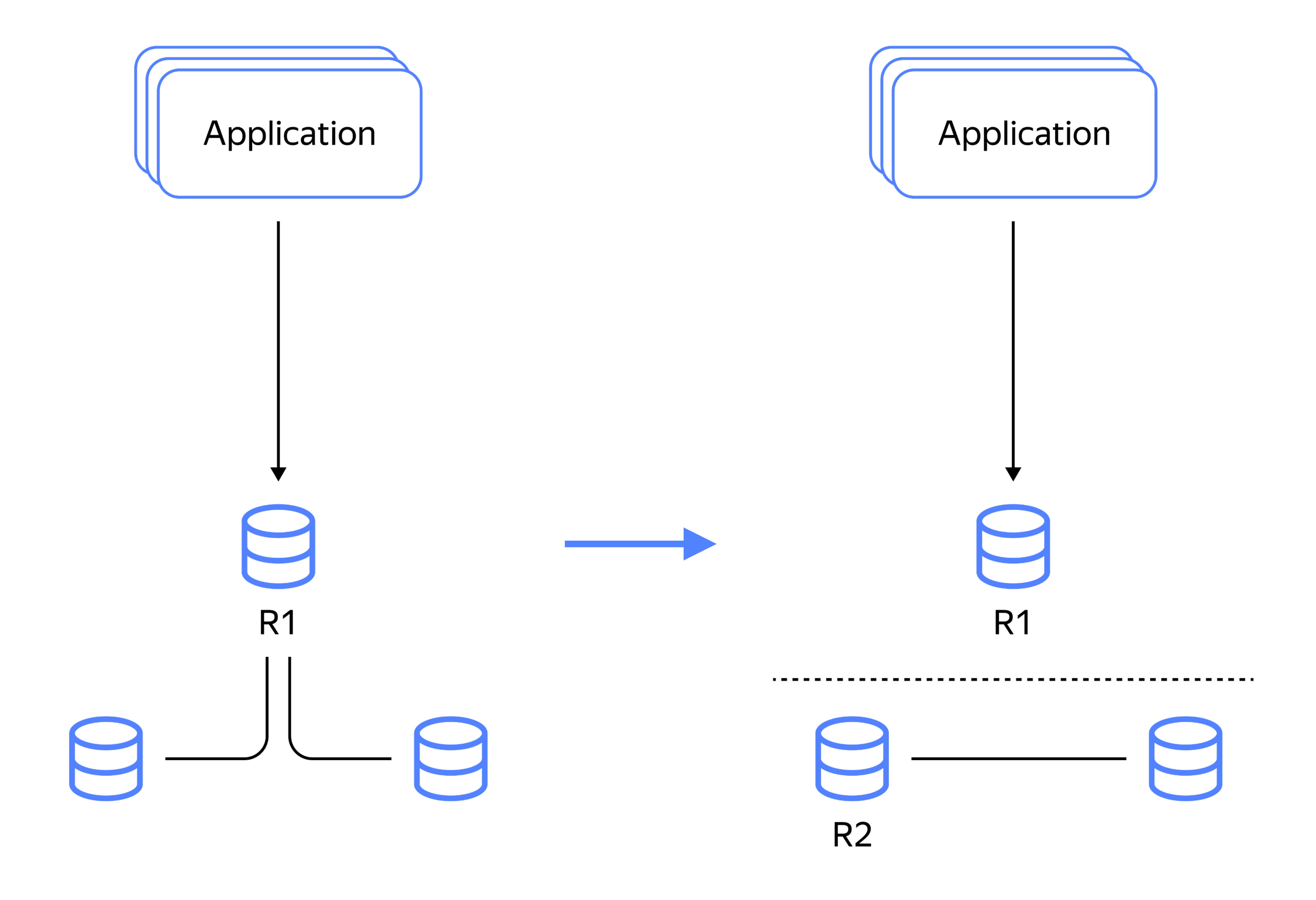
Sentinel

- Partitioned primary не закрывается от нагрузки

Cluster

- Partitioned primary не закрывается от нагрузки

Пример проблемы



Проблемы

Sentinel

- Partitioned primary не закрывается от нагрузки
- Проблемы с изоляцией

Cluster

- Partitioned primary не закрывается от нагрузки

Проблемы

Sentinel

- Partitioned primary не закрывается от нагрузки
- Проблемы с изоляцией
- Совмещение Sentinel и Redis на одних узлах может порождать неотказоустойчивые конфигурации

  Например, две ноды

Cluster

- Partitioned primary не закрывается от нагрузки

Проблемы

Sentinel

- Partitioned primary не закрывается от нагрузки
- Проблемы с изоляцией
- Совмещение Sentinel и Redis на одних узлах может порождать неотказоустойчивые конфигурации

  Например, две ноды

Cluster

- Partitioned primary не закрывается от нагрузки
- Голосуют только primary
  → количество шардов имеет значение

Проблемы

Sentinel

- Partitioned primary не закрывается от нагрузки
- Проблемы с изоляцией
- Совмещение Sentinel и Redis на одних узлах может порождать неотказоустойчивые конфигурации

  Например, две ноды

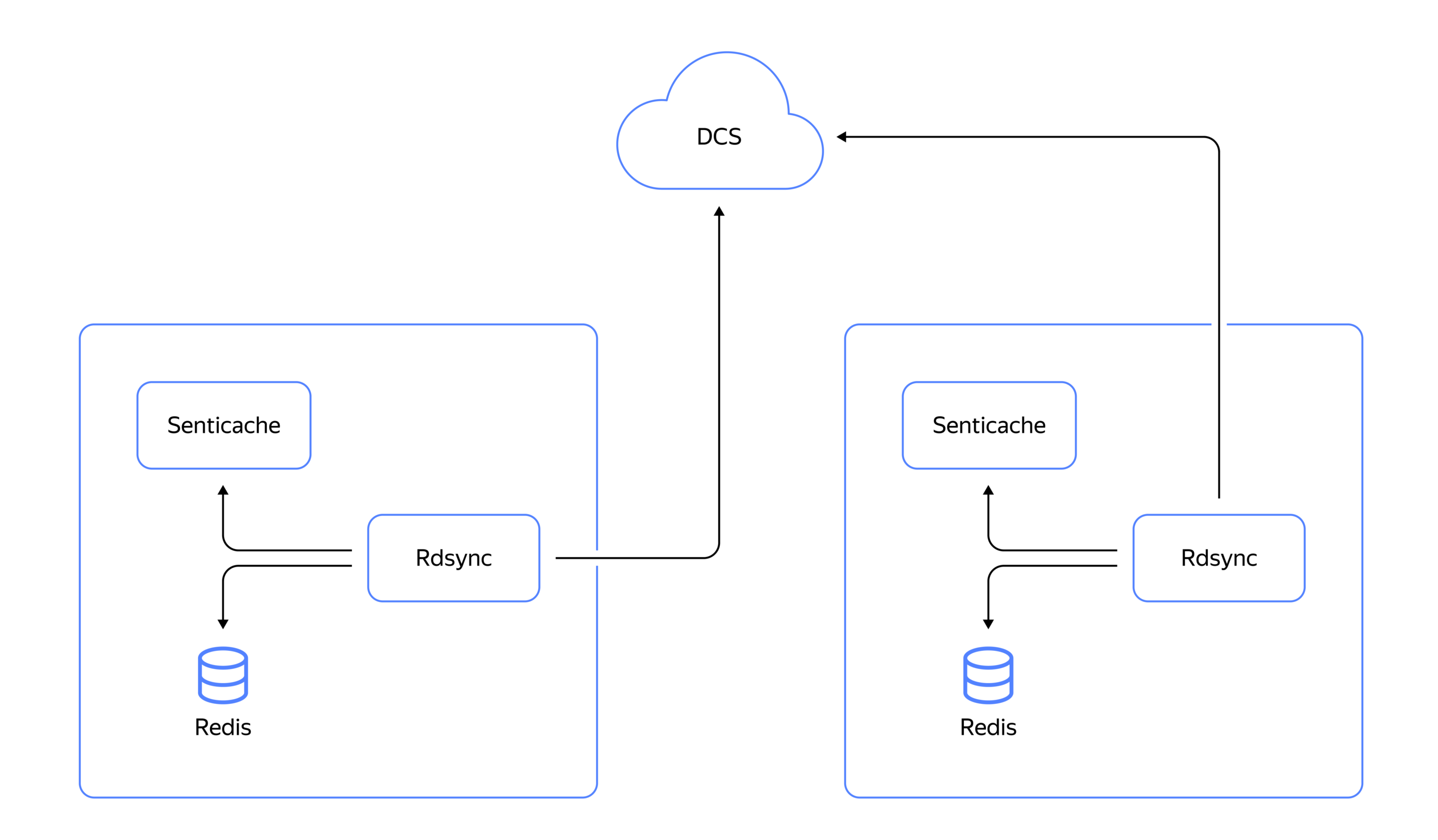
Cluster

- Partitioned primary не закрывается от нагрузки
- Голосуют только primary → количество шардов имеет значение
- Распределение primary по доменам отказа тоже

Быстрые решения

- Закрываем partitioned primary
  goo.su/MykxpP
- Совмещаем Sentinel и Redis на одном наборе хостов и пишем в документации, какие конфигурации отказоустойчивы
- Запрещаем создавать кластеры меньше чем с тремя шардами
- Следим за распределением primary по доменам отказа и размазываем их

Альтернатива — rdsync



Роли инстансов rdsync

Manager

- Держит лок в DCS
- Следит за живостью нод, принимает решение о Failover
- Делает Switchover
- Плюс всё, что делает Candidate

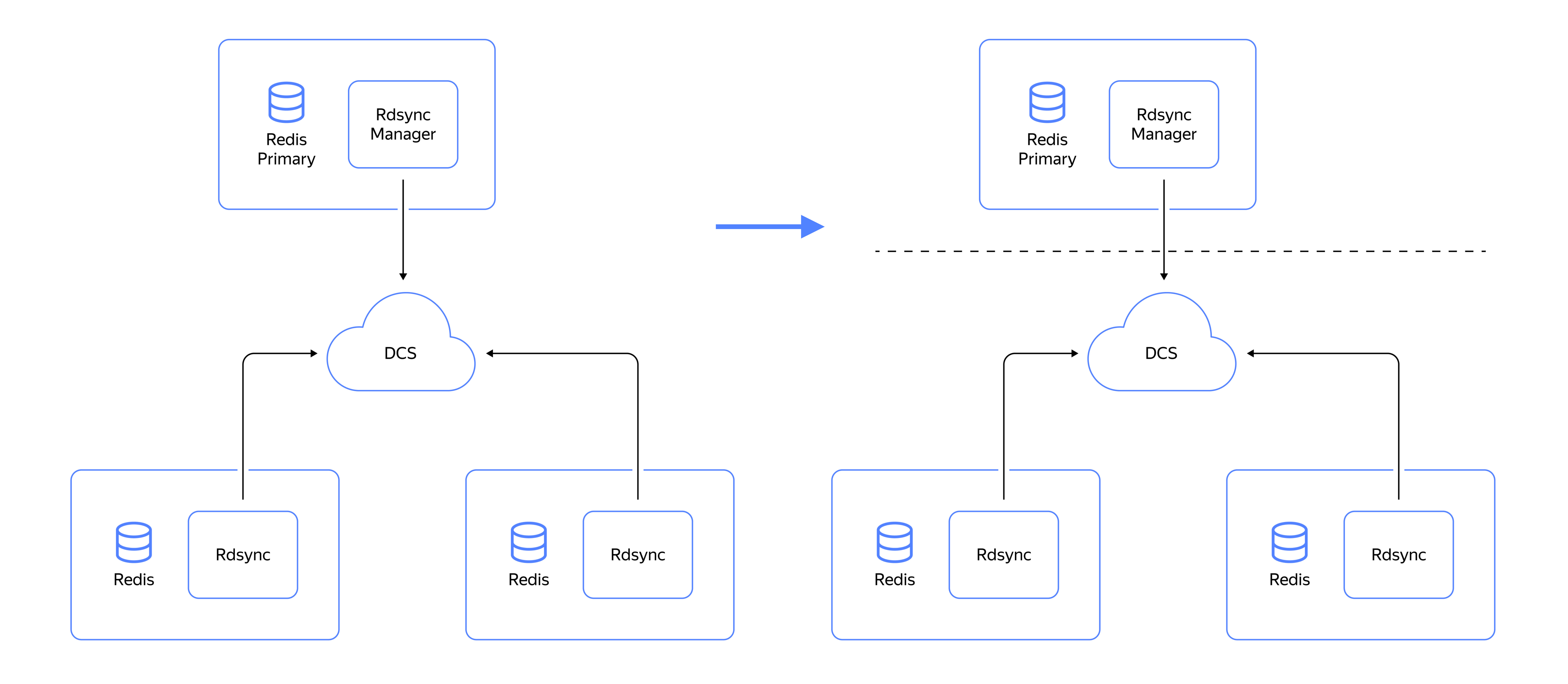
Candidate

- Пытается взять лок и стать Manager
- Обновляет информацию о локальном Redis в DCS
- Обновляет состояние в локальном Senticache
- Закрывает локальный Primary при потере связи с DCS

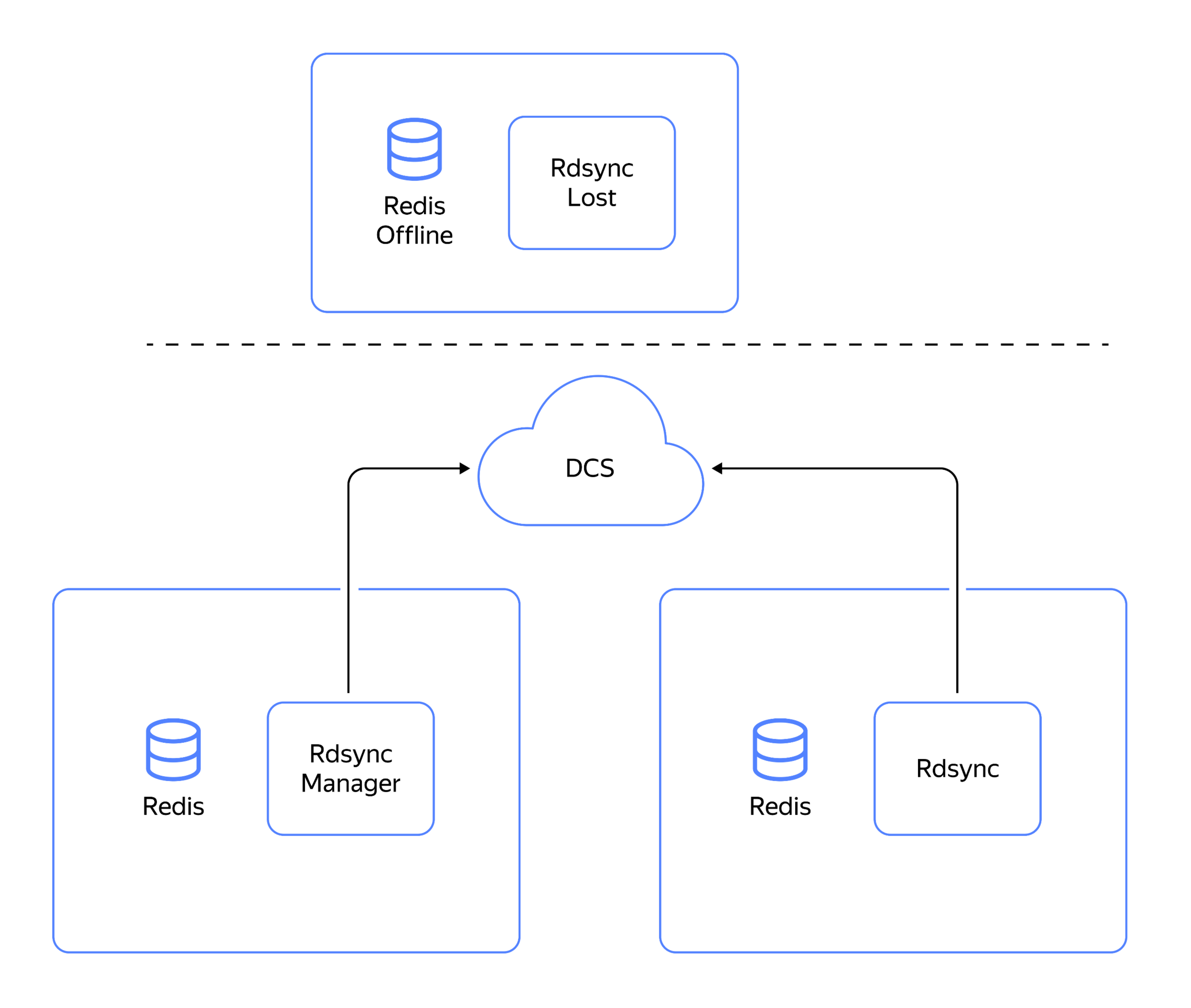
Senticache

- Форкаем Sentinel
- Отрываем всю логику про failover и соединения вовне
- Добавляем отдельную команду для обновления внутреннего состояния

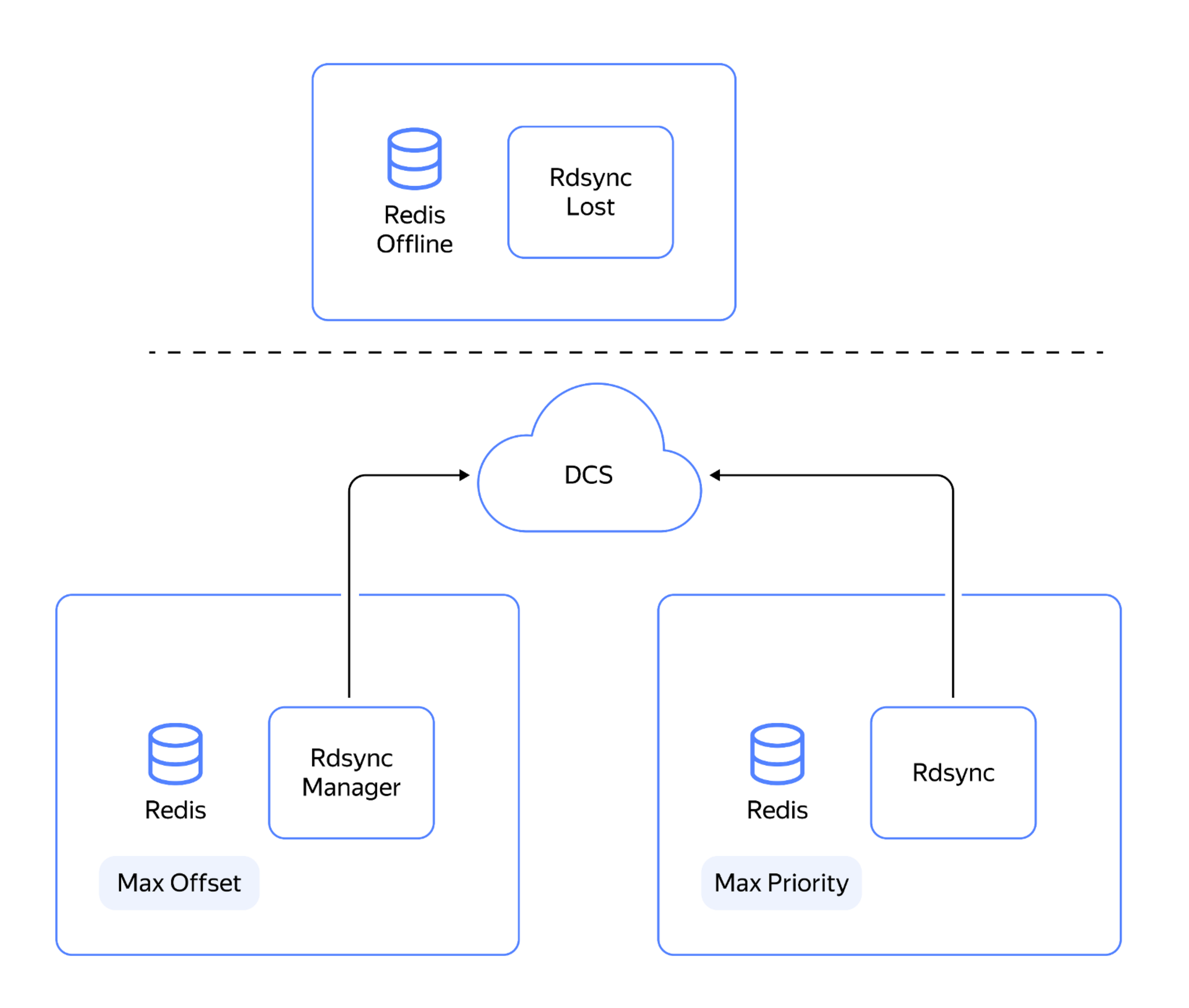
Failover: начальная конфигурация



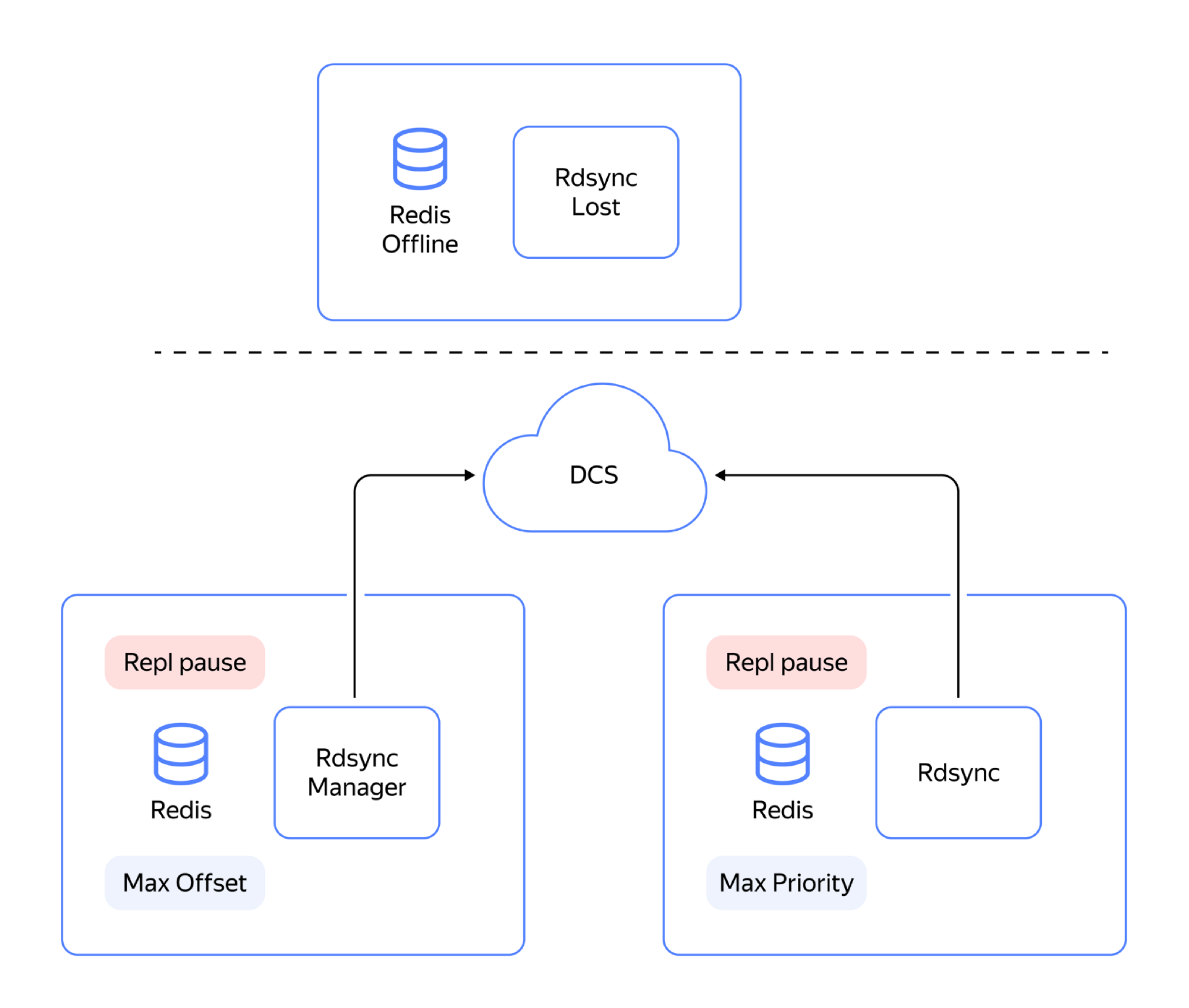
Failover: новый manager + offline* старого primary



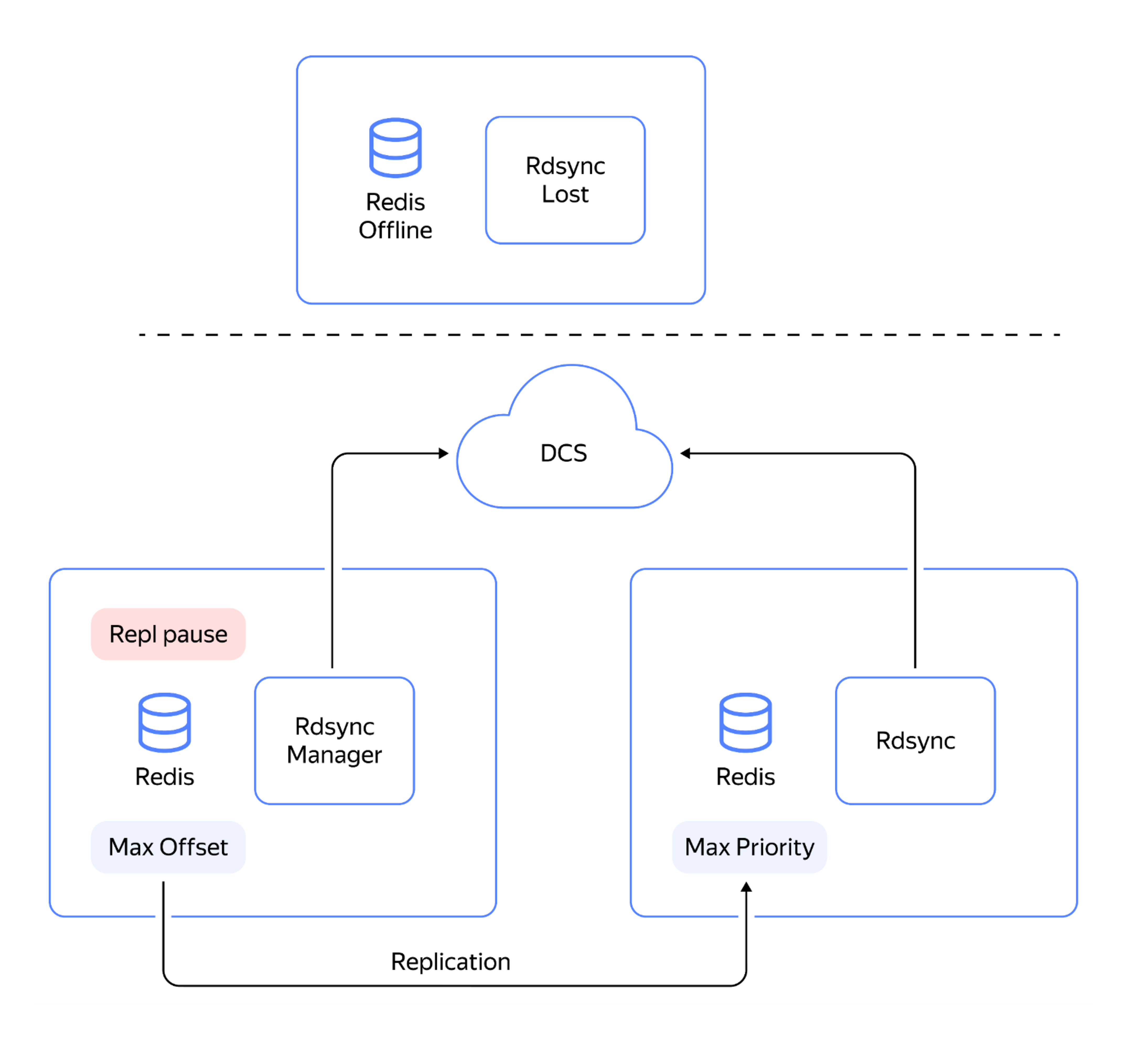
Failover: новый manager выясняет текущее состояние



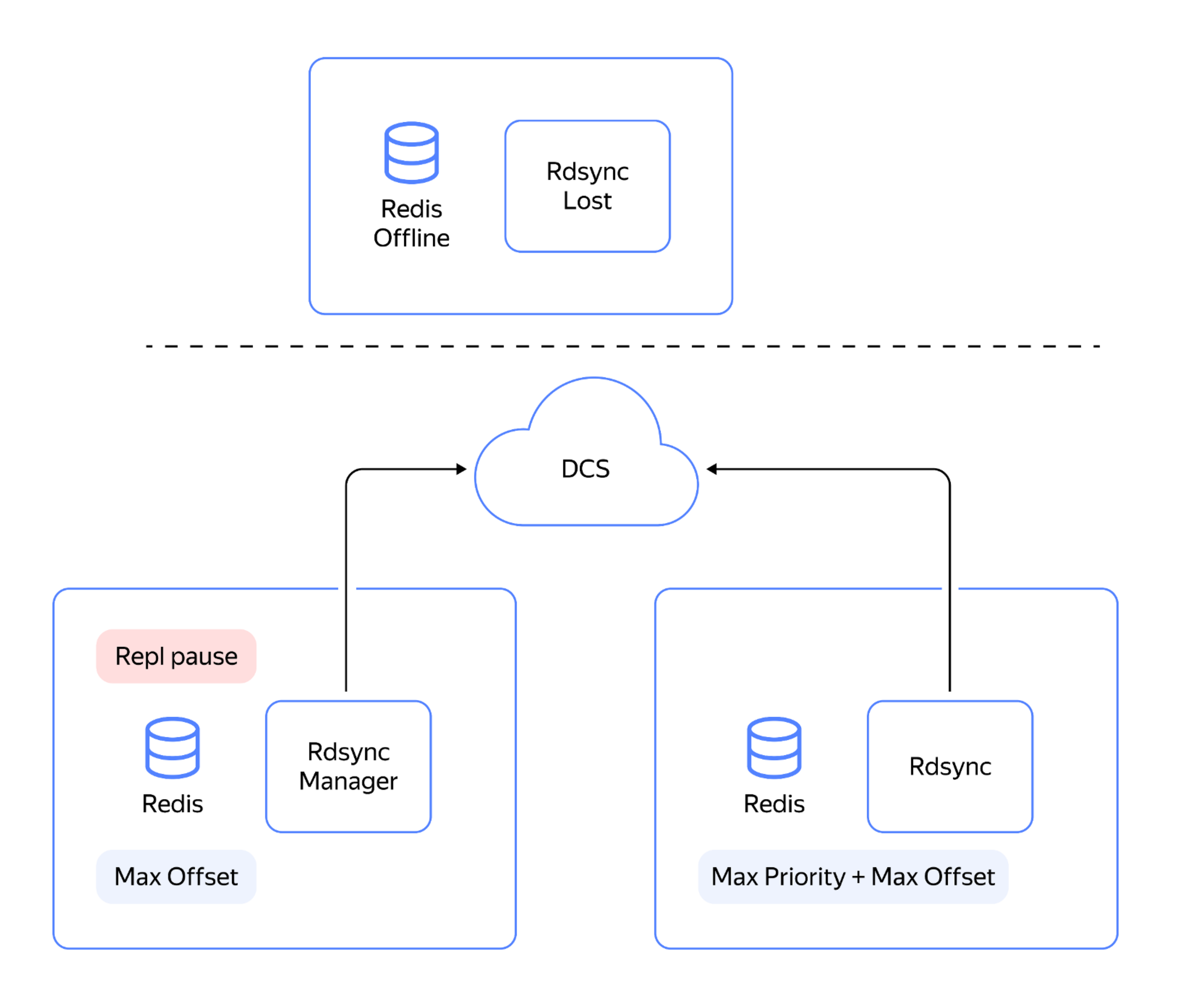
Failover: остановка репликации*



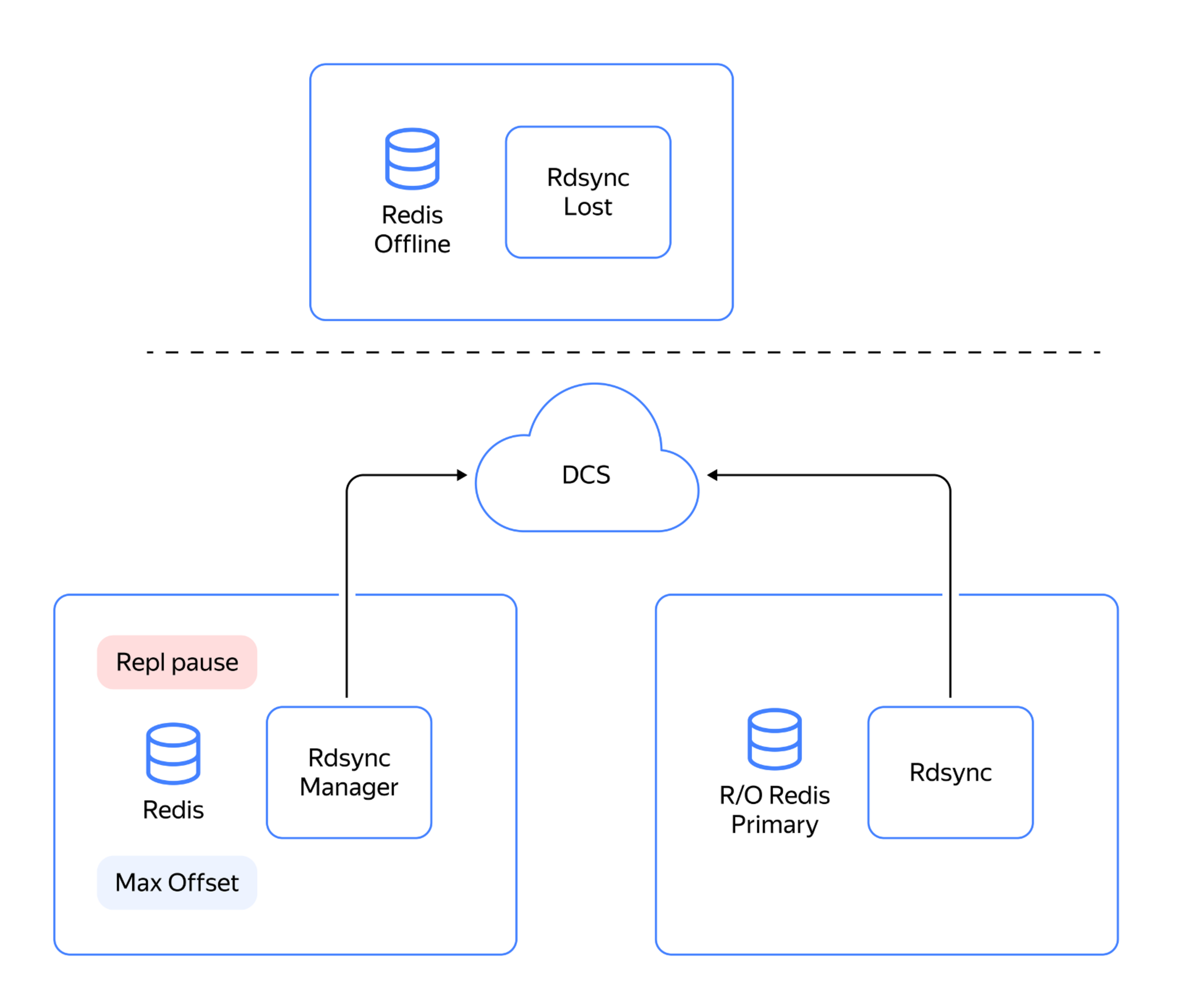
Failover: догоняем самую приоритетную до самой свежей*



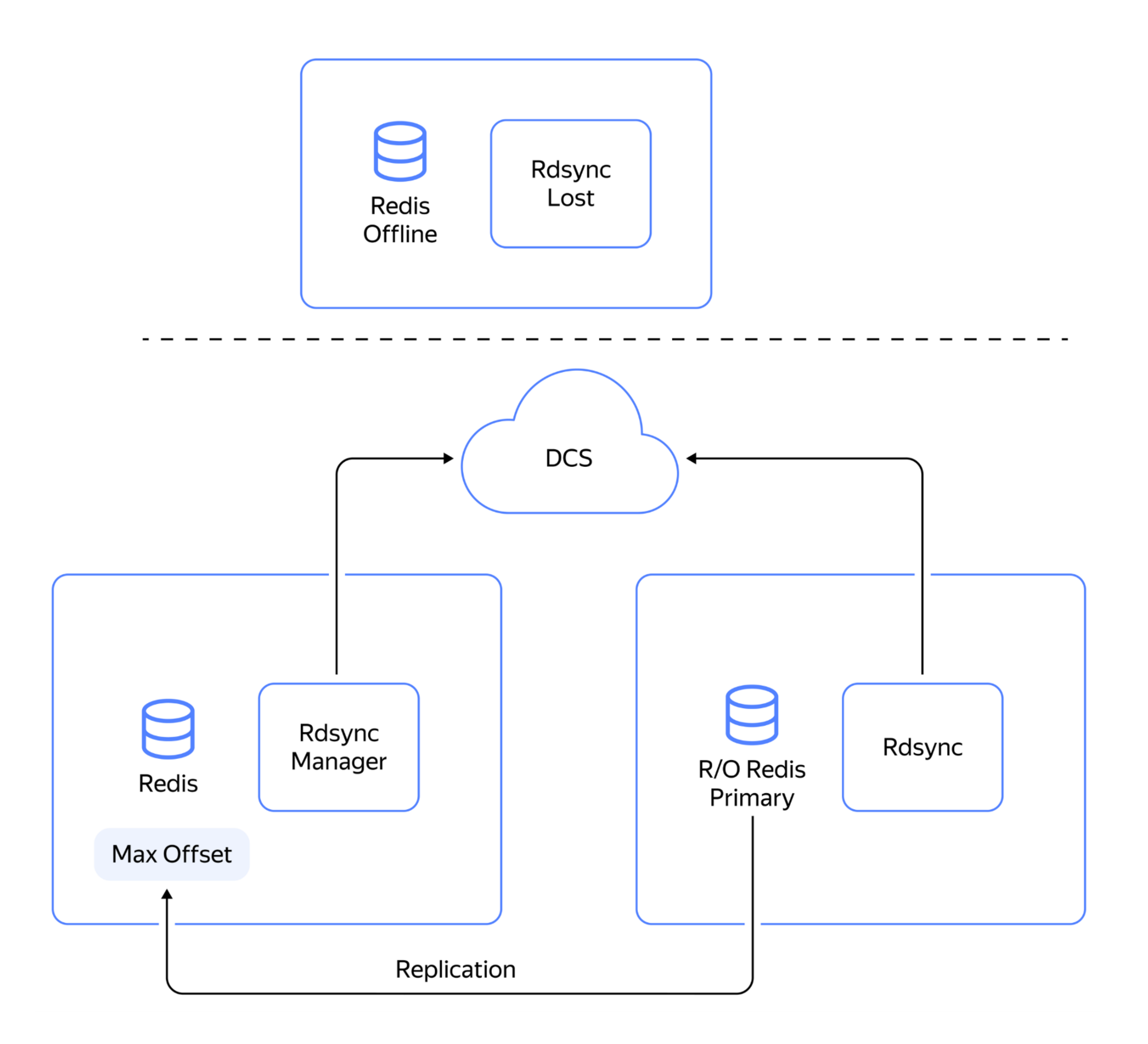
Failover: догоняем самую приоритетную до самой свежей*



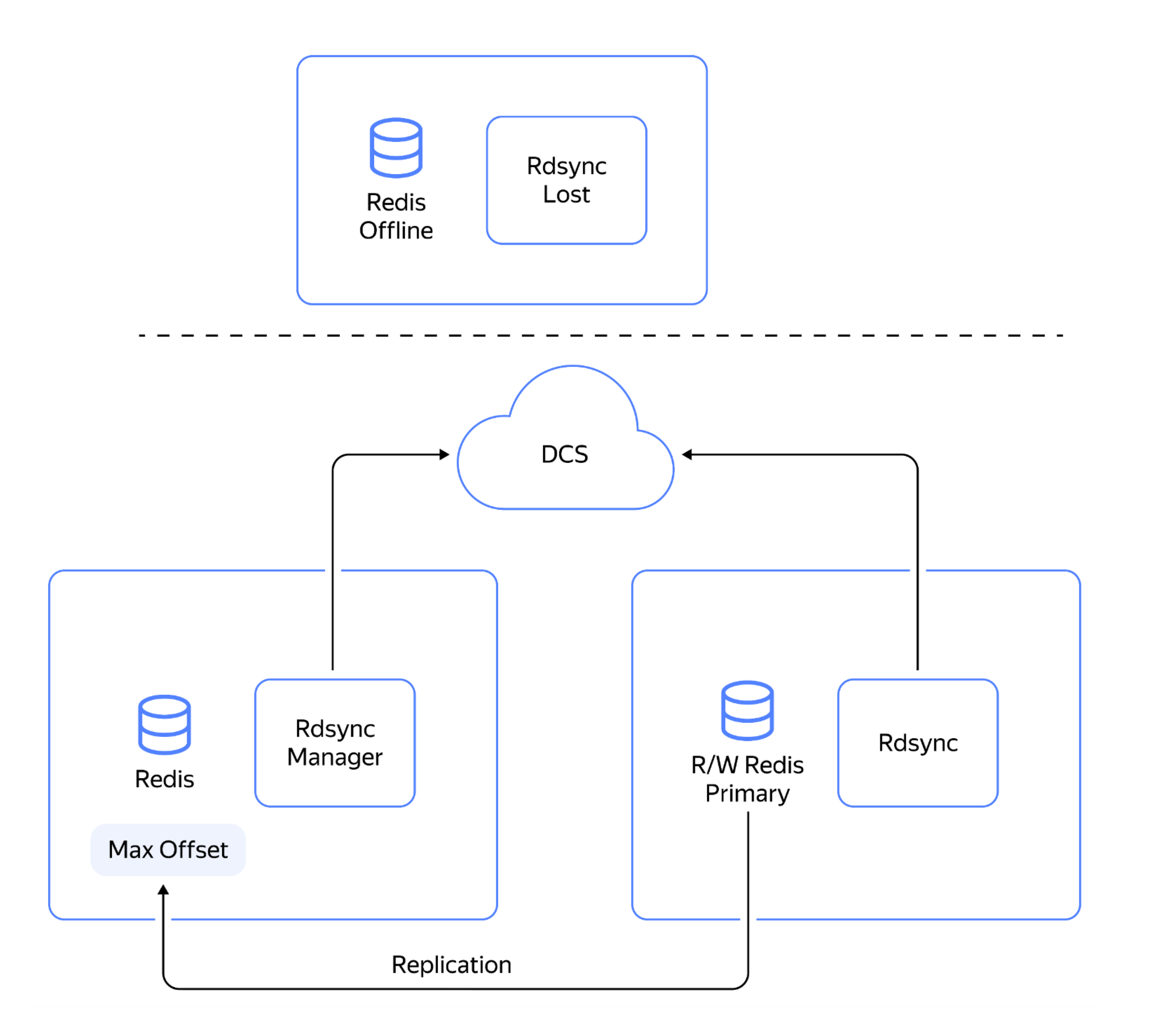
Failover: promote



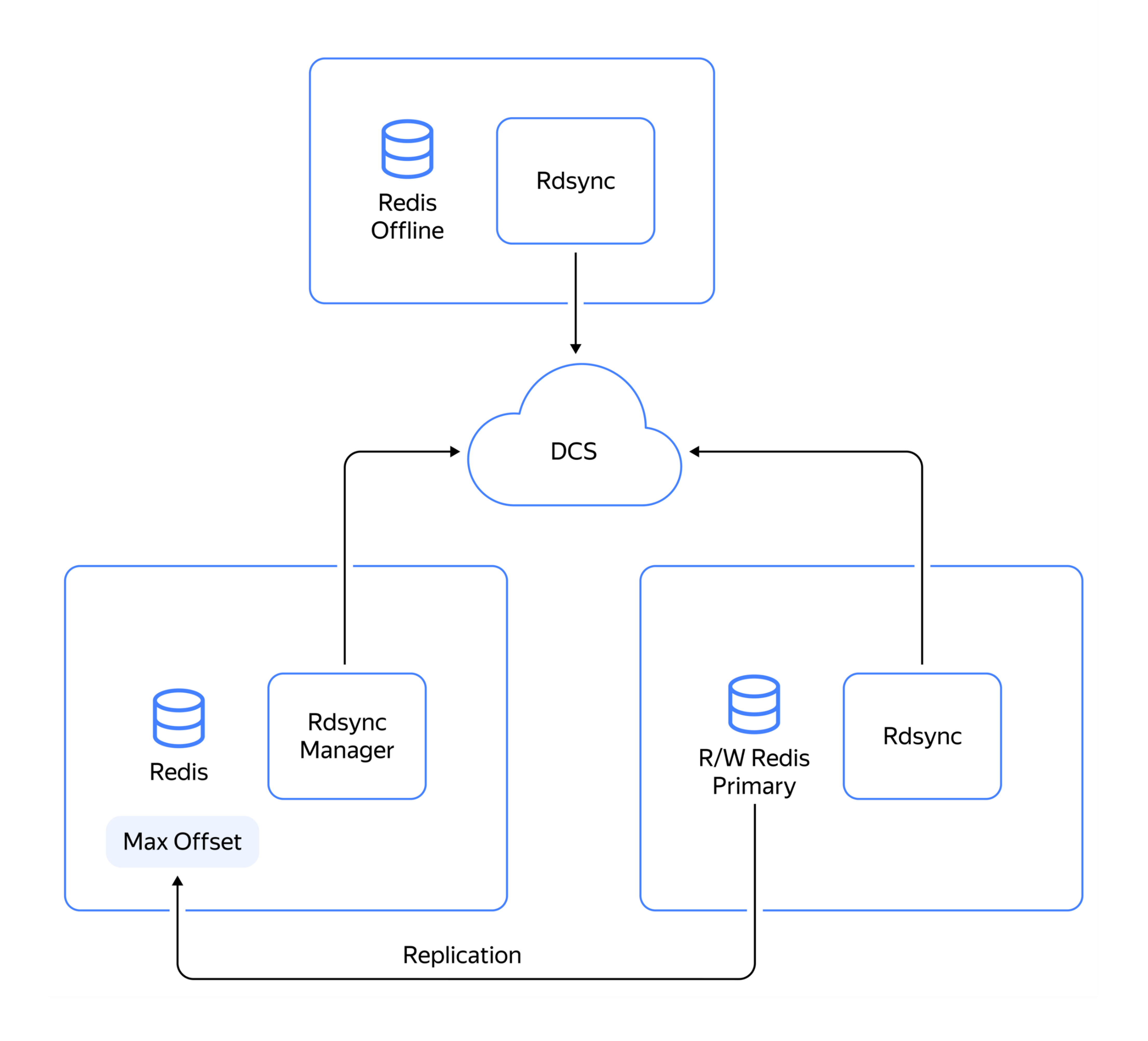
Failover: поворачиваем реплики



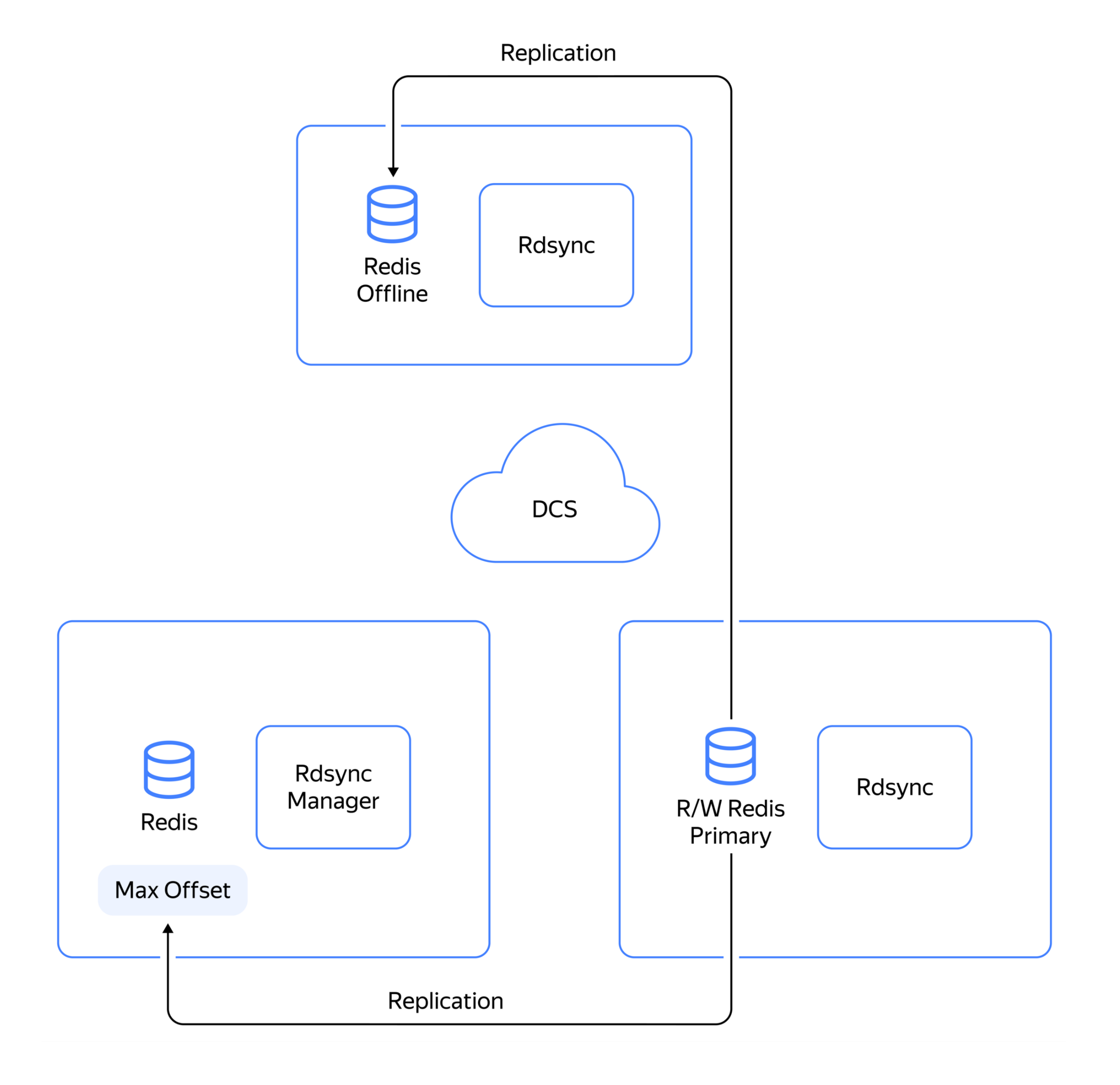
Failover: открываемся на запись



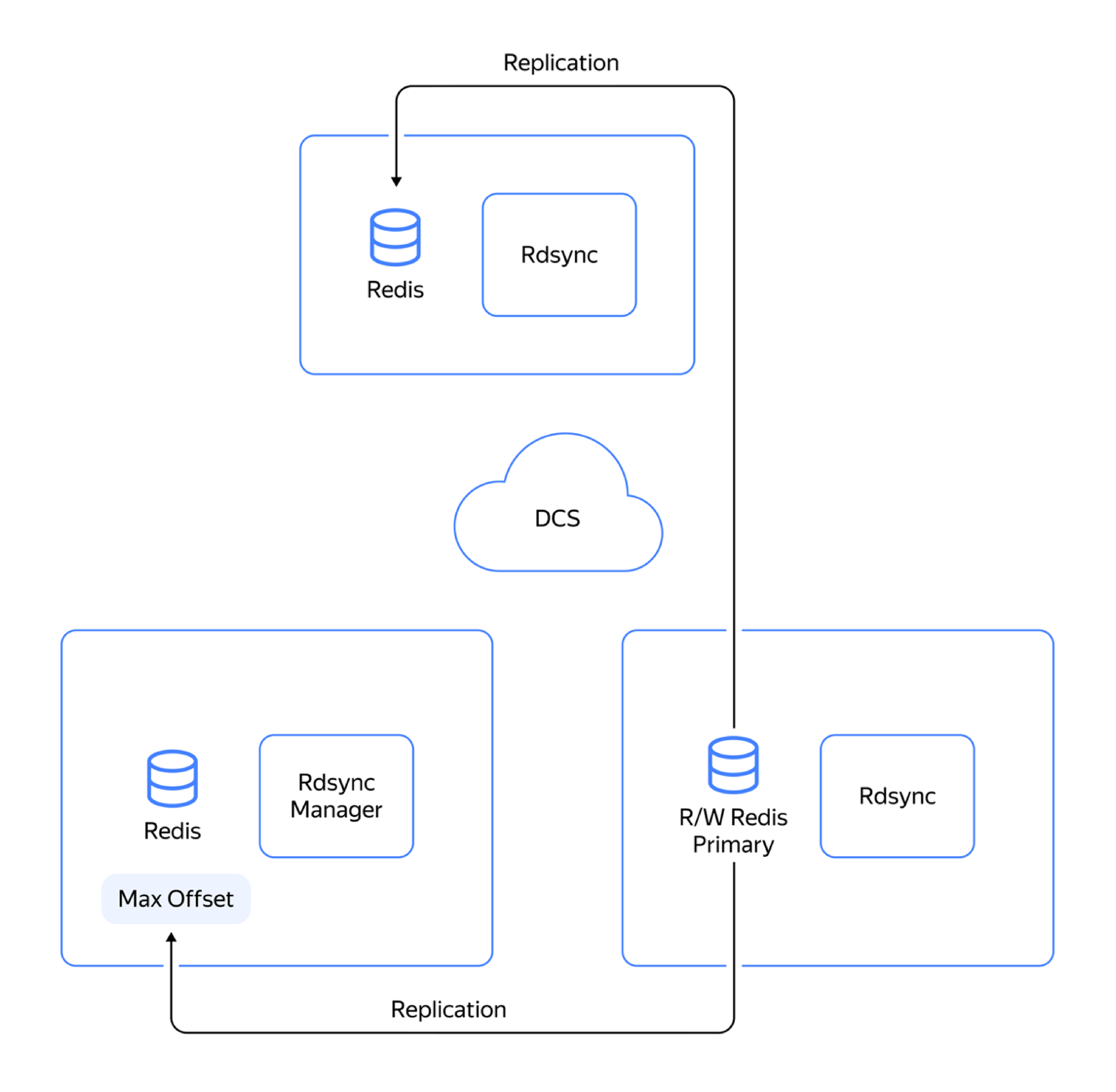
Failover: partition healing



Failover: переналиваем старый primary



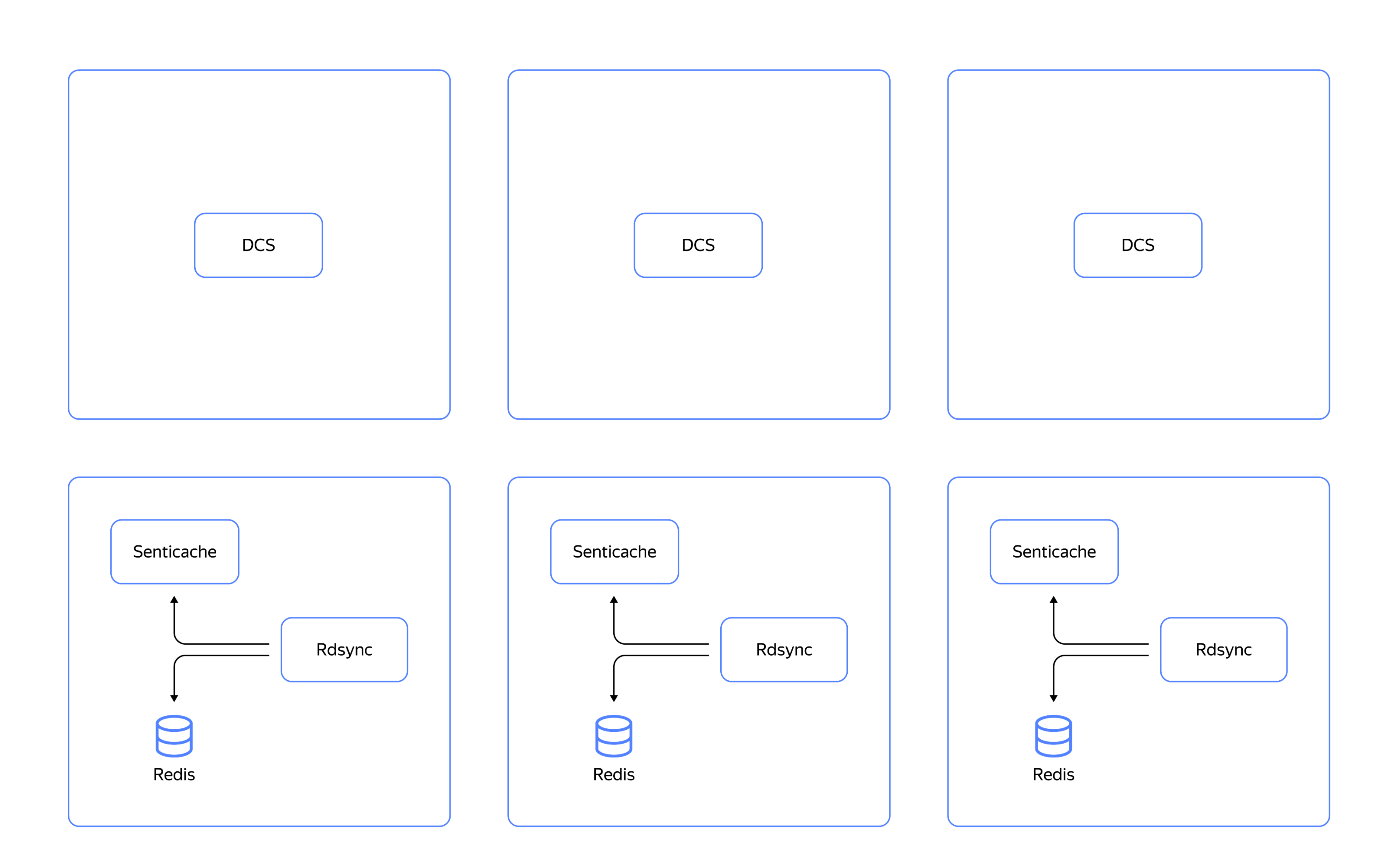
Failover: открываем его на чтение



Тесты

Sentinel

- Функциональные на сценарии failover/switchover
- Jepsen
  Только network partition



Итого

1

Chubby-way rocks

(https://goo.su/pAMv)

2

Патчить Redis не так уж страшно

3

Используйте Jepsen(-like)-тесты, если не хотите терять данные

(https://jepsen.io)

Спасибо!

Евгений Дюков

Разработчик Managed Databases
secwall@yandex-team.ru



Яндекс

Обратная связь
и комментарии
к докладу по ссылке





HighLoad++ 2022

Яндекс